CFA/VISHNO 2016

Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation

R. Nicol[a], L. Gros[a], C. Colomes[b], E. Roncière[c] et J.-C. Messonnier[d]
[a]Orange Labs, 2 Av Pierre Marzin, 22307 Lannion, France
[b]Orange labs, rue du Clos Courtel, 35510 Cesson Sévigné, France
[c]Radio France, 116 Avenue du Président Kennedy, 75220 Paris Cedex 16, France
[d]CNSMDP, 209 Avenue Jean Jaurès, 75019 Paris, France
rozenn.nicol@orange.com

Les contenus audio voient aujourd'hui l'émergence de multiples formats multicanaux (5.1, 7.1, 10.2, 22.2, Auro 3D, Dolby Atmos) destinés à renforcer l'immersion sonore. Mais ces formats nécessitent de s'équiper de systèmes de reproduction impliquant un grand nombre de haut-parleurs, ce qui freine leur diffusion. La technologie binaurale offre une solution élégante en proposant une écoute sur casque basée sur des haut-parleurs virtuels (processus de "binauralisation"). C'est justement l'objectif du projet FUI BiLi. Un des axes de recherche porte sur l'évaluation de la perception du rendu binaural associé. Différents aspects de la chaîne de reproduction binaurale sont considérés. L'étude présentée ici porte sur l'impact du premier maillon : la prise de son qui va constituer la matière première de l'information spatiale portée aux oreilles de l'auditeur. Dans ce but, une expérience a été menée afin de comparer le rendu, à l'issue de la binauralisation, d'enregistrements d'une pièce radiophonique par différentes techniques de prise de son spatialisée, incluant des couples stéréophoniques, des arbres multicanaux (5.0 ou 8.0), des microphones Ambisonics (ordre 1 et ordre 4), et des têtes artificielles. Deux expériences ont été mises en place. Dans la première, le sujet doit reporter graphiquement la position et les trajectoires des principales sources sonores. La seconde consiste en un jugement de préférence par paire. Dans les deux cas, l'expérience est assortie d'un questionnaire pour faire le lien avec les attributs perceptifs. Les résultats indiquent que les jugements des sujets experts sont sensiblement plus discriminants que ceux des sujets naïfs. La spatialisation artificielle (basée sur des microphones d'appoint) et un système multicanal 5.0 se distinguent dans le premier test par la précision de leur localisation. En revanche, dans le second test, une préférence marquée pour la tête KU100 de Neumann est observée. L'étude ouvre des questions sur les aspects méthodologiques.

1 Introduction

Depuis quelques années, on assiste à une multiplication des formats audio qui cherchent à améliorer la qualité d'expérience des contenus audio ou audiovisuels. Ces nouveaux formats misent sur un accroissement du nombre de canaux principalement pour enrichir l'information spatiale et contribuer à améliorer l'immersion sonore. Ainsi on est passé d'une reproduction focalisée sur l'espace frontal (stéréophonie) à des systèmes 2D pour une spatialisation horizontale ("son surround" 5.1, 7.1), puis à des systèmes 3D (formats 10.2, 22.2, Auro 3D, Dolby Atmos) qui permettent de reproduire plus ou moins partiellement une information d'élévation. Cependant, pour bénéficier de ces nouveaux formats, il faut disposer d'un équipement multi haut-parleurs à la fois onéreux et encombrant. Ces équipements sont rarement compatibles avec les contraintes du grand public. Or, si les producteurs de contenus (radios et télévisions par exemple) veulent promouvoir ces nouveaux formats audio multicanaux, il faut justifier que les contenus associés soient effectivement écoutables par une audience large, ce qui n'est pas le cas aujourd"hui. Il existe néanmoins une solution à ce problème : il s'agit du traitement de binauralisation qui permet d'adapter un signal multicanal à une écoute sur casque et qui est basé sur la création de haut-parleurs virtuels par synthèse binaurale. La binauralisation des contenus audio multicanaux est au cœur du projet FUI BiLi qui a pour objectif de proposer des solutions à la fois pour le grand public et les professionnels de l'audio, avec l'ambition d'améliorer la qualité d'immersion par la personnalisation du traitement (www.bili-project.org). Un premier volet du projet BiLi porte sur le développement d'outils de binauralisation personnalisée. Mais il importe aussi d'évaluer la perception du résultat final. C'est de cette question que traite l'étude présentée ici et qui a été menée dans le cadre des travaux sur l'évaluation de la Qualité d'Expérience (QoE) du projet BiLi [1]. La perception d'un contenu binauralisé ne dépend pas que des performances de la binauralisation, mais aussi de la chaîne allant de la création à la reproduction de ce contenu. Dans ce qui suit, on s'intéresse au premier maillon de cette chaîne que constitue la prise de son.

L'article décrit une expérience dans laquelle on évalue la perception du même contenu sonore enregistré par différents systèmes de prise de son incluant des couples stéréophoniques, des arbres multicanaux (5.0 ou 8.0), des microphones Ambisonics (ordre 1 et ordre 4), et des têtes artificielles. Dans tous les cas, le contenu est reproduit sur un casque, moyennant une étape de binauralisation si nécessaire. L'objectif est de comparer le résultat perçu en fonction du système de prise de son. L'expérience comprend deux tests d'écoute : un premier test où le sujet reporte la position et les éventuels mouvements des principales sources sonores, et un second test où des jugements de préférence sont collectés, la prise de son par la tête artificielle Neumann KU100 étant choisie comme référence. Dans la première partie, la constitution des stimuli audio utilisés pour l'expérience est décrite, en précisant les procédures de prise de son et de postproduction des contenus. La seconde partie présente le protocole expérimental. Les résultats sont analysés et discutés dans la troisième partie. La dernière partie conclut par une discussion.

Figure 1 – Prise de vue d'une session d'enregistrement.

2 Stimuli

Les stimuli audio de l'expérience sont extraits d'une session d'enregistrement d'une dramatique radiophonique ("Deux femmes pour un fantôme" de René de Obaldia) au

CNSMDP (Conservatoire National Supérieur de Musique et de Danse de Paris) [2][3] . Cette pièce a été choisie pour la variété de contenus (parole, chant, bruitages et musique) qu'elle offre. Mais surtout, la liberté de mouvement du personnage du fantôme ouvre la possibilité d'explorer l'espace et son rendu, ce qui répond à nos préoccupations. La salle d'enregistrement est le Grand Plateau d'Orchestre du CNSMDP, qui permet de mettre en scène l'ensemble de ces situations et mouvements. En complément des sources naturelles (acteurs, bruiteur, instruments de musique), un dispositif multi haut-parleurs a été mis en place pour enregistrer le générique de fin. Cette séquence permet de disposer d'un contenu pour lequel la position des sources sonores est parfaitement identifiée et reproductible d'une séance d'enregistrement à l'autre. Elle est constituée de 13 sources artificielles qui sont distribuées sur l'ensemble de la sphère 3D autour de l'auditeur. La totalité de la séquence dure une quinzaine de minutes.

Au total, 11 systèmes de prise de son spatialisée ont été utilisés (voir Tableau 1). Les principales catégories y sont représentées : couples stéréophoniques, arbres multicanaux 4.0, 5.0 ou 8.0, microphones Ambisonics d'ordre 1 et 4, têtes artificielles. En complément, les comédiens sont équipés de micros sans fil, et des microphones d'appoints sont disposés dans la salle pour certains évenements ponctuels (musiciens, bruitage, ...). Pour des raisons évidentes, les 12 systèmes n'ont pu être placés simultanément au même endroit pendant la même session d'enregistrement (Voir Figure 1). Il a été décidé de choisir la tête artificielle KU100 comme prise de son de référence. Le monitoring du réalisateur a ainsi été basé sur cette prise de son. Elle est donc présente à toutes les sessions d'enregistrement. Par suite, la tête artificielle KU100 définit aussi le centre du repère. Pour une session donnée, deux ou trois systèmes de prise de son lui sont associés. Ils sont disposés au plus proche de la tête KU100, en cherchant à minimiser les interactions entre les différents dispositifs. Il a fallu au total six sessions pour collecter les enregistrements correspondants à l'ensemble des systèmes de prise de son. Cette procédure séquencielle d'enregistrement pose la question de la reproductibilité du contenu d'une session à l'autre. Une reproductibilité parfaite est impossible, mais la variabilité est minimisée par le fait que tous les intervenants (acteurs, musiciens, bruiteur) sont professionnels et sont rompus à ce genre d'exercice.

L'étape de postproduction a principalement consisté à appliquer le traitement de binauralisation. Seuls les enregistrements des 3 têtes artificielles n'ont pas été modifiés. Dans le processus de binauralisation, le choix des HRTF a une influence fondamentale, notamment selon que le jeu de HRTF correspond ou non aux HRTF individuelles de l'auditeur. Mais étudier ce paramètre aurait multiplié les conditions expérimentales. Pour cette première expérience, il a donc été décidé de fixer ce paramètre et de considérer exclusivement le cas de HRTF non indivividuelles, ce qui, aujourd'hui, constitue d'ailleurs la condition la plus probable dans le contexte de diffusions grand public. C'est l'ingénieur du son en charge de la postproduction qui a sélectionné le jeu de HRTF qui lui semblait le plus approprié, avec la consigne de choisir le jeu qui permettait d'obtenir le meilleur rendu pour chaque système de prise de son, au sens de ses critères personnels d'ingénieur du son. Au final, à l'exception des prises de son Ambisonics, c'est le même jeu de HRTF (jeu référencé 1040 de la base Listen [4]) qui a été utilisé pour tous les enregistrements. En complément des 11 enregistrements, deux autres versions ont été créées à la postproduction : il s'agit des versions référencées JML+A et SA dans le Tableau 1, ce qui nous amène à un total de 13 versions à comparer.

Tableau 1 – Liste des systèmes de prise de son [2].

Type	Nom	Descriptif
Binaural natif	KU100	tête artificielle (Neumann)
	TH	tête artificielle conçue par L. Hô (DPA 4060)
	TL	tête artificielle conçue par B. Lagnel (DPA 4060)
Couple stéréophonique	AB	couple AB (Neumann TLM 50)
Arbres multi-canaux	IRT	arbre 4.0 (Schoeps MK5)
	JML	arbre 5.0 (DP4 4041)
	JML+A	mixage de l'arbre JML et des microphones d'appoints
	5100	arbre DPA 5100 (5.0)
	MMAD5	arbre 5.0 (Schoeps MK21)
	MMAD8	arbre 8.0 (Schoeps MK21)
Microphones Ambisonics	SF	Soundfield® (ordre 1)
	EM	Eigenmike® (ord. 4)
Spatialisation artificielle	SA	synthèse binaurale à partir des microphones d'appoints

3 Protocoles expérimentaux

L'objectif de l'étude est d'évaluer et de comparer la reproduction binaurale de différents formats de prise de son spatialisée (Tableau 1). Il faut bien garder à l'esprit qu'il ne s'agit pas de comparer la perception avant et après binauralisation, ce qui reviendrait à évaluer les artefacts de la binauralisation. Ici on se place dans le cas exclusif d'une reproduction binaurale et on souhaite comparer la robustesse de différents systèmes de prise de son au regard du traitement de binauralisation. Dans les critères évalués, on s'intéresse principalement à la question de la reproduction de l'information spatiale. Cependant, il ne faut pas perdre de vue que la perception d'un système de reproduction sonore, quel qu'il soit, est multi-dimensionnelle [1][5]. Si les attributs liés à la spatialisation sont de première importance, il ne faut pas perdre de vue les autres dimensions (timbre notamment). Une des difficultés de cette étude a été de construire un protocole capable de rendre compte, de façon plus ou moins complète, de ces différentes dimensions, tout

Tableau 2 – Liste des extraits utilisés pour les stimuli.

Nom	Durée	Description
Extrait 1	19s	Déplacement de l'acteur autour de l'auditeur (parole et bruitage)
Extrait 2	26s	musique (flûte traversière, contrebasse, percussions)
Extrait 3	16s	dialogue entre les deux actrices
Extrait 4	24s	générique sur ambiance de musique

en préservant une charge expérimentale "raisonnable". Pour résoudre cette difficulté, l'expérience a été décomposée en deux tests basés sur deux protocoles totalement distincts. Le premier test consiste principalement en un jugement de localisation, le second sur un jugement de préférence. Par ailleurs, une autre question d'intérêt porte sur l'effet de l'expertise des participants sur leurs jugements. Différentes catégories d'auditeurs sont donc considérées : des auditeurs "non experts" (*i.e.* dénués d'une quelconque expérience, ou sensibilisation, de l'écoute des sons), des auditeurs expérimentés (*i.e.* habitués à l'écoute et l'évaluation de sons et notamment de sons spatialisés), des auditeurs "experts" (*i.e.* qui travaillent dans le domaine du traitement et de la spatialisation des sons). Les paradigmes expérimentaux sont détaillés dans les paragraphes qui suivent. Les deux tests se sont déroulés dans des salles différentes, mais chaque fois dans un environnement calme (salle de test acoustiquement traitée et isolée). Les équipements audio étaient identiques. Les stimuli étaient diffusés à travers un casque Seinnheiser HD650 (Amplificateur de casque TASCAM MH-40 MK II), depuis un ordinateur équipé d'un carte Digigram carte audio VX 222 et d'un convertisseur Numérique Analogique externe 3Dlabs DAC 2000 24bits.

3.1 Test 1 [6]

Dans ce premier test, on demande à l'auditeur d'identifier les principales sources sonores et d'indiquer leur position et leur(s) éventuel(s) déplacement(s). On fait l'hypothèse que la fiabilité de l'information spatiale capturée par un système de prise de son donné impacte la précision du jugement de localisation. En évaluant ce dernier, on a donc une mesure de la capacité d'un système à retranscrire l'information spatiale. Le report de ces informations se fait sous forme graphique sur une feuille de papier calque posée sur une feuille de papier millimétré, au centre de laquelle la tête de l'auditeur est figurée de façon symbolique (voir Figure 2). Cette tâche de description de la scène sonore est réalisée pour un stimulus unique d'une durée égale à 1min40s. Ainsi l'auditeur effectue sa tâche de localisation en continu. Le stimulus est constitué de la concaténation de 4 extraits de la séquence de la dramatique "Deux femmes pour un fantôme" enregistrée au CNSMDP (voir Tableau 2). Un silence de 5 secondes est inséré entre chaque extrait. Les extraits ont été choisis pour obtenir une certaine variété de contenus : voix, musique, sources en mouvement et sources statiques. Est inclus un extrait du générique, dans lequel les sources

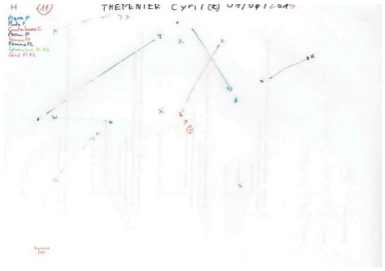

Figure 2 – Exemple de description de la scéne sonore par un auditeur dans le test 1 (échelle : 1cm pour 1m).

correspondent à des haut-parleurs alimentés par des signaux de parole. Cet extrait se limite à quatre sources combinant deux locuteurs masculins et deux locuteurs féminins. En complément, et à l'issue de l'écoute de la séquence associée à chaque système de prise de son, chaque participant répond à un questionnaire simple comportant quatre questions portant sur le réalisme, la précision spatiale, la sensation d'espace, et la distance maximale ressentie. Pour les 3 premières questions, la réponse se fait sur une échelle de 5 degrés de 0 à 4.

Douze auditeurs ont pris part au test : six auditeurs experts et six non experts. Pour chaque sujet, le test se déroulait en commençant par l'écoute de l'intégralité de la séquence afin qu'il identifie l'ensemble des sources sonores. Pour cette écoute, la séquence utilisée était choisie aléatoirement parmi les 13 prises de son. Ensuite l'auditeur écoutait successivement les 13 versions de la séquence associées aux 13 prises de son, selon un ordre aléatoire différent pour chaque participant. Chaque version n'était écoutée qu'une fois. Pendant l'écoute, le sujet reportait les sources sonores sur le graphique (voir Figure 2) au fur et à mesure des évenements sonores. A la fin de chaque version, il complètait le questionnaire avant de passer à la version suivante. Au total, le test pour un seul participant durait de l'ordre de 1h30min. Deux ou trois pauses étaient imposées.

3.2 Test 2 [7]

Ce second test vise à évaluer le degré de préférence des différentes captations par rapport à une prise de son de référence. La version enregistrée par la tête artificielle KU100 a été choisie comme la version de référence. Ce choix, partiellement arbitraire, a été motivé par le fait qu'il s'agit d'une version au format dit binaural natif, correspondant à une prise de son par une tête artificielle, par opposition aux versions issues des enregistrements par le couple stéréophonique, les arbres multicanaux, les microphones Ambisonics ou les microphones d'appoints, dans lesquelles les signaux binauraux sont obtenus par synthèse binaurale qui présente un certain nombre d'artefacts en comparaison d'une prise de son naturelle. Le second argument de ce choix provient de la session d'enregistrement, pour laquelle la captation par la KU a été définie comme écoute de référence pour le réalisateur.

Pour évaluer ce degré de préférence, le paradigme expérimental repose sur un test de comparaison par paire, chaque paire étant composée de deux versions d'une même séquence sonore, une version qui correspond à l'une des douze captations testées et l'autre à la prise de son binaurale par la KU100. Pour une paire donnée, il est demandé au

sujet d'évaluer son niveau de préférence de la version "A" par rapport à la version "REF" sur une échelle de 7 degrés. Chaque paire est présentée deux fois mais dans un ordre différent : pour une des présentations, la KU100 est associée à REF, tandis que pour l'autre, la KU100 est associée à A. L'interface de test est basée sur une adaptation de l'interface du logiciel CRC-SEAQ utilisé pour les tests MUSHRA et BS1116 [8][9]. Le logiciel ne permettant pas de proposer des échelles catégorielles, il est précisé, dans la consigne, que l'échelle doit être considérée comme une échelle de catégorie. Chaque dispositif de captation est ainsi évalué pour trois séquences sonores courtes (une vingtaine de secondes) qui correspondent aux extraits 1, 2 et 4 utilisés pour le test 1 (voir Tableau 2). Un total de 72 paires (3 séquences x 12 paires x 2 présentations) a été évalué.

En complément, il était demandé au sujet de répondre à un questionnaire qui propose une liste de 10 attributs perceptifs pouvant être associés à la perception de sons binauraux [5] : Crédibilité/ Réalisme, Immersion sonore, Ampleur de la scène sonore, Equilibre spatial, Externalisation, Précision spatiale, Coloration, Transparence/ Respect du timbre, Effet de salle/ Réverbération, Relief. Pour chaque attribut, le sujet doit indiquer si la définition proposée lui semble compréhensible (2 réponses possibles : oui ou non), et s'il a utilisé cet attribut pour effectuer son jugement de préférence (3 réponses possibles : non, oui - pour quelques paires, oui - pour toutes les paires).

Au total, 24 sujets ont participé au test : 8 sujets non experts, 8 sujets expérimentés, 8 sujets experts. Pour la moitié des sujets, la liste d'attributs et leur définition sont données avant le test, afin d'évaluer si cette information améliore l'acuité du jugement. Le test débute par une séance d'apprentissage formée de 3 paires permettant aux participants de se familiariser avec l'interface et avec les différences qu'ils auront à juger pendant le test. Lors de la phase d'apprentissage, l'auditeur est invité à régler le niveau sonore, qui ne devra plus être modifié lors du test. Les 72 paires sont ensuite évaluées. Il était recommandé de bien écouter chaque séquence une première fois complètement. Il était également précisé qu'il est possible de passer librement d'une version à l'autre (A ou REF) à tout moment. Enfin le sujet était invité à faire des pauses si nécessaire.

4 Résultats

4.1 Test 1

Pour chaque participant, 13 graphiques décrivant la scène sonore sont collectés pour les différents systèmes de prise de son. Il a fallu construire une méthode d'analyse spécifique à ce protocole expérimental. Un graphique de référence basé sur les positions physiques des sources sonores relevées lors de la session d'enregistrement a été établi. Les graphiques produits par les sujets ont été comparés à ce dessin de référence afin d'évaluer dans quelle mesure les positions et mouvements des sources sonores restitués par chaque système de captation étaient conformes à la réalité physique. Cette comparaison s'est focalisée sur une sélection de sources : voix de l'acteur pour l'extrait 1, flûte, contrebasse, percussions pour l'extrait 2, haut-parleurs de l'extrait 4. Dans chaque cas on s'intéresse séparément aux informations de déplacement, de direction et de distance. La comparaison visuelle du dessin de référence et du dessin reproduit par l'auditeur permet d'identifier si l'information spatiale est conforme à la réalité, si elle présente un léger biais ou si elle n'est pas conforme. Ainsi, pour un système de prise de son donné, on comptabilise, pour la totalité des sources considérées et la totalité des participants, le pourcentage de localisations conformes et de localisations non conformes. Le ratio de ces deux pourcentages définit le critère de classement des systèmes de captation : un ratio supérieur à 1 signifie que le nombre de localisations conformes est supérieure aux localisations non conformes. Plus ce ratio est élevé, meilleure est la localisation. On observe des différences sensibles entre les jugements de localisation des sujets experts et ceux des sujets non experts. Ces derniers ont en général des performances moindres de localisation, ce qui tend à réduire la valeur du ratio et peut affecter le classement relatif des systèmes. Le Tableau 3 présente le classement des différents systèmes de prise de son selon ce critère. Le classement fait apparaître un relatif consensus entre experts et non experts pour le système SA, l'e système JML+A, les têtes artificielles TH et TL, dont la localisation semble la plus précise, ainsi que pour le microphone Soundfield®, la tête artificielle KU100, les arbres MMAD5 et MMAD8 dont la localisation au contraire s'avère la moins précise. Cependant, alors que le microphone Soundfield® est classé sensiblement de la même façon par les experts et non experts, seuls les experts jugent meilleure la localisation du microphone Eigenmike®. De même, les classements

Tableau 3 – Classement des systèmes de prise de son sur la base de l'analyse des reports graphiques des jugements de localisation du test 1.

Rang	Experts	Non experts
1	SA (37%)	SA (32%)
2	TH (38%)	TL (33%)
3	IRT (36%)	JML+A (31%)
4	JML+A (40%)	TH (34%)
5	TL (33%)	5100 (31%)
6	MMAD8 (37%)	AB (29%)
7	EM (33%)	MMAD5 (23%)
8	5100 (31%)	KU100 (28%)
9	MMAD5 (32%)	IRT (26%)
10	SF (31%)	MMAD8 (21%)
11	JML (24%)	SF (23%)
12	KU100 (28%)	JML (17%)
13	AB (27%)	EM (17%)

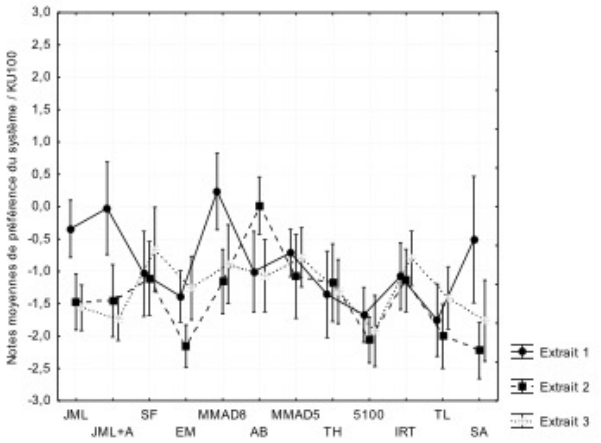

FIGURE 3 – Notes moyennes de préférence (référence : KU100) en fonction de l'extrait sonore.

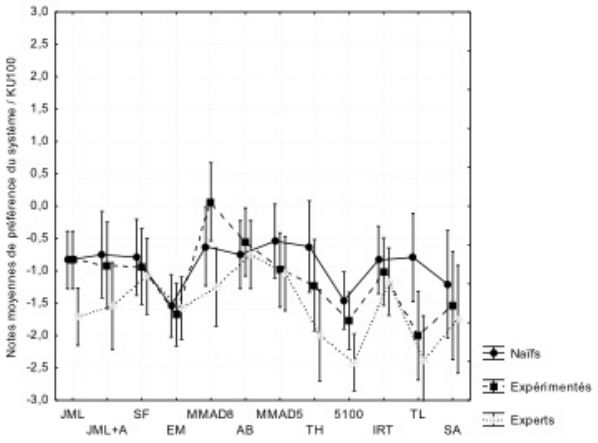

FIGURE 4 – Notes moyennes de préférence (référence : KU100) en fonction de l'expertise.

experts et non experts divergent fortement pour l'arbre IRT et le couple AB. Alors que l'arbre IRT est bien classé (rang 3) par les experts, il est classé au rang 9 par les non experts. L'inverse se produit pour le couple AB. Pour ce dernier, il faut rappeler que la stéréophonie est devenue une sorte de référence d'écoute, ce qui peut expliquer qu'elle soit bien interprétée par des auditeurs non experts.

4.2 Test 2

Etant donné que, dans le Test 2, chaque paire de stimuli a fait l'objet d'une double présentation, on analyse d'abord la cohérence des jugements entre ces deux présentations. Globalement, les auditeurs sont cohérents dans leurs jugements. Sur les vingt-quatre auditeurs, seuls trois auditeurs obtiennent un taux de cohérence inférieur à 70%. Les experts semblent plus cohérents avec des taux exclusivement supérieurs à 80% (taux moyen des experts : 89%). Les auditeurs expérimentés et non experts recueillent quant à eux des taux moyens respectifs de 82% et 73%. Les taux de cohérence sont également analysés en fonction du système de captation et de l'extrait. L'effet du système de captation s'avère prédominant. Ainsi le couple AB génère plus d'incohérences dans les jugements de préférence (taux = 67%) que le système DPA5100 (92%) ou le microphone Eigenmike® (90%). Cette tendance pourrait suggérer que l'auditeur est partagé entre sa préférence pour la référence culturelle que constitue la stéréophonie et sa préférence esthétique.

Une fois menée l'analyse des cohérences, une note de préférence est calculée en moyennant les préférences obtenues pour les deux présentations A-REF et REF-A. La Figure 3 montre les notes moyennes de préférence recueillies pour les différents systèmes de captation, en comparaison avec la KU100, et en fonction de l'extrait sonore. Il apparaît qu'aucun système de captation n'est préféré à la tête artificielle KU100 choisie comme système de captation de référence dans ce test (moyennes inférieures ou égales à zéro). Avec les extraits 2 et 3, c'est même la tête KU100 qui est clairement préférée, en moyenne, et ce pour la majorité des systèmes (excepté pour le couple AB avec l'extrait 2). Avec l'extrait 1, la KU100 est également préférée, sauf pour les captations JML et JML+A, l'arbre MMAD8 et, de façon moins unanime, pour le système SA.

TABLEAU 4 – Résultats de l'ANOVA.

Effet	F	p
Expertise	5,46	,012*
Système	7,44	,000*
Système*Expertise	1,5	0,075
Extrait	11,59	,000*
Extrait*Expertise	1,35	0,267
Système*Extrait	4,14	,000*
Système*Extrait*Expertise	1,47	,030*

La Figure 4 illustre l'effet potentiel de l'expertise et montre un renforcement de la préférence pour la KU100 avec le niveau d'expertise, sauf pour les systèmes SF, EM, AB et IRT. Afin de tester la significativité de cet effet, une analyse de variance ANOVA est réalisée sur les notes individuelles de préférence en considérant les facteurs Expertise (à trois niveaux), Systèmes (à 12 niveaux) et Extraits (à 3 niveaux). L'effet de l'expertise s'avère assez faible et peu significatif (voir Tableau 4).

Enfin l'analyse des réponses au questionnaire montre que, dans l'ensemble, les définitions proposées pour chacun des 10 attributs semblent bien comprises. Le critère qui recueille le plus d'incompréhensions, mais de façon limitée, est l'externalisation (3 non experts et 2 expérimentés), même après l'expérience potentielle de cette sensation (voir Section 3.2). L'ampleur de la scène sonore, la coloration, l'effet de salle et l'équilibre spatial sont non compris de façon sporadique (1 ou 2 incompréhensions, par des auditeurs différents - non experts, expérimentés ou même experts- et quel que soit le protocole). Bien que tous compris, les attributs ne sont pas tous utilisés de la même façon. La Figure 5 montre les effectifs obtenus pour chacun des attributs dans les différentes catégories d'utilisation proposées aux auditeurs. Pour un peu plus de la moitié des attributs (Ampleur de la scène sonore, Coloration, Réalisme,

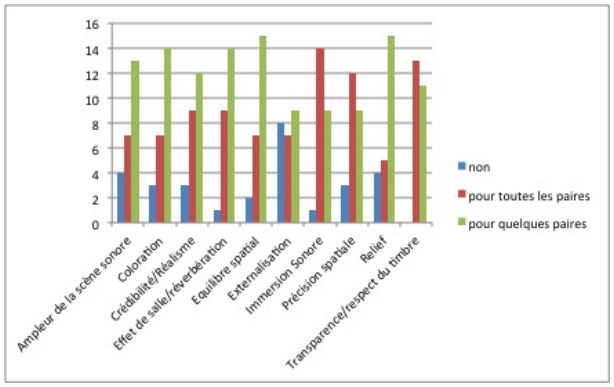

FIGURE 5 – Effectifs d'utilisation des attributs.

Effet de salle, Equilibre spatial et Relief), on obtient un profil d'utilisation similaire : *i.e.* la majorité des auditeurs les utilise pour quelques paires, puis moins de testeurs pour toutes les paires et enfin quelques testeurs ne les utilisent pas. Trois attributs semblent davantage utilisés systématiquement : l'immersion sonore, la précision spatiale et enfin le respect du timbre. L'externalisation est l'attribut qui recueille le plus de "non utilisé", en nombre sensiblement équivalent aux deux autres catégories d'utilisation. La différence d'effectifs avec les autres attributs dans cette catégorie est cohérente avec le nombre d'incompréhensions obtenu pour la définition de ce critère. La présentation des définitions avant ou après passation du test ne modifie pas ou peu ces constatations. Il semble également que les experts aient davantage varié les attributs utilisés, en comparaison des auditeurs expérimentés et non experts.

5 Discussion

Il semble difficile d'établir un classement des systèmes en termes de préférence sur la base du test 2, d'une part car les différences observées entre systèmes ne semblent pas souvent significatives, et d'autre part car ces différences dépendent assez fortement de l'extrait écouté. En revanche, il apparaît clairement que la tête artificielle KU100 est le système de captation préféré. Ce résultat en soi est en contradiction avec les conclusions du test 1, dans lequel le système KU100 obtient des scores médiocres et est classé respectivement au rang 12 et 8, par les auditeurs experts et non experts, au regard du critère de localisation (voir Tableau 3). Cette contradiction suggère que le ressenti global des scènes sonores ne dépend pas de la performance des systèmes à reproduire les sons selon leur emplacement dans la scène enregistrée/réelle, performance qui est en revanche évaluée dans le test 1. Dans le test 2, l'analyse des réponses au questionnaire sur les attributs utilisés pour le jugement de préférence fait ressortir les attributs relatifs à l'immersion sonore, la précision spatiale et au respect du timbre, comme les plus utilisés. Les résultats de cette étude montrent donc que le choix d'une solution technologique de restitution audio spatialisée va fortement dépendre du but recherché et de l'application visée, selon que l'on souhaite reproduire précisément les sons aux positions prévues ou réelles dans le cas d'un enregistrement (jeu vidéo par exemple), ou que l'on recherche davantage une reproduction qui privilégie l'immersion, le timbre, la précision. D'un point de vue méthodologique, le test 2 prouve qu'il est possible d'effectuer une tâche de jugement de préférence entre deux versions « spatialisées » d'une scène sonore. Ce jugement de préférence permet d'éviter l'utilisation d'échelles et d'attributs qui restent encore à définir pour caractériser la perception multi-dimensionnelle de scènes sonores spatialisées. Le test 1 a, quant à lui, montré la capacité des auditeurs à positionner, et ce de façon dynamique, des sources sonores sur une feuille de papier. Il serait intéressant de comparer les résultats de localisation obtenus avec cette procédure avec ceux obtenus via des tests de localisation plus classiques utilisant, le plus souvent des sons de courte durée et sans mouvement. Si des résultats comparables étaient obtenus, la procédure « par report graphique, en continu, pour une scène sonore complète » pourrait apporter à la fois des informations sur la localisation tout en étant compatible avec une mesure plus écologique du ressenti. Evidemment, ces conclusions ne valent que pour les conditions expérimentales testées correspondant à une reproduction exclusivement sur casque, pour laquelle le traitement de binauralisation repose sur un choix arbitraire de HRTF non individuelles. De nouvelles expériences sont nécessaires pour étudier le lien entre agrément et attributs perceptifs dans le cas d'une reproduction sur réseau multi hauts-parleurs. L'influence du choix des HRTF est aussi à évaluer, en utilisant soit des HRTF individuelles, soit des HRTF personnalisées.

Références

[1] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. F. G. Katz, L. S. R. Simon, A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering, Proc. of the EAA Joint Symposium on Auralization and Ambisonics, (2014).

[2] E. Roncière, Compte-rendu de captation (Rapport interne du Projet BiLi) : "Fiction CNSMDP : 2 femmes pour un fantôme de René de Obaldia", (2014).

[3] www.bili-project.org/comparaison-de-systemes-de-prise-de-son-3d-resultats

[4] recherche.ircam.fr/equipes/salles/listen/

[5] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, S. Weinzierl, A Spatial Audio Quality Inventory (SAQI), Acta Acustica united with Acustica, 100(5), pp. 984-994 (2014).

[6] C. Colomes, Compte-rendu des résultats de test de méthodologie d'évaluation de différents dispositifs de captation (Rapport interne du Projet BiLi) - Test 1, (2015).

[7] L. Gros, Compte-rendu des résultats de test d'évaluation de différents dispositifs de captation (Rapport interne du Projet BiLi) - Test 2 : Evaluation des préférences par rapport à la KU100, (2015).

[8] ITU-R BS.1534-3 : Method for the subjective assessment of intermediate quality level of audio systems (October 2015).

[9] ITU-R BS.1116-1 : Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems (October 1997).